Probabilistic Multi-Person Tracking with Relative Position Measurements

Kanji TANAKA and Mamoru MINAMI
Department of Human and Artificial Intelligence System, Faculty of Engineering,
University of Fukui
3-9-1, Bunkyo, Fukui, 910-8507,
JAPAN
http:// *rc.his.fukui-u.ac.jp/*

Abstract: **This paper presents a particle filtering framework for tracking multiple persons with a monocular camera. So far, most of techniques based on particle filtering rely on an assumption that measurements on pose and speed of moving persons are sufficiently precise. Unfortunately, such an assumption is often violated due to measurement noises as well as irregular movements of persons. To deal with the problem, we have developed a technique for measuring relative position between persons using occlusion reasoning. In particular, we show how the use of relative position measurements can improve the tracking performance, even in difficult situations where two persons frequently overlap in images.**

Key-Words: **visual tracking, occlusion reasoning, particle filter**

1. INTRODUCTION

This paper is concerned with a problem of visual tracking of multiple persons by a fixed monocular camera in cluttered office environments. The problem of visual tracking is a crucial problem for robot vision, motion analysis as well as situation understanding. In this domain, the technique of particle filtering recently attracts much interests as it provides a consistent scheme for tracking the pose and the speed of targets in a general multi-modal probabilistic framework [1][2].

So far, most of techniques based on particle filtering rely on an assumption that measurements on the pose and the speed of moving objects are sufficiently precise [3][4]. Firstly, they estimate the 3D pose of individual objects from raw or background-subtracted images by using a pre-learned or pre-defined relationship between the camera and the world coordinate systems. In many cases, the depth information cannot be acquired by a monocular camera. So, some a priori knowledge on such as the pose of floor where the person is walking on, as well as the shape of person etc are typically used [3][5]. Secondly, they associate each object with a target being tracked by using some models of appearances and/or motions often based on the above introduced assumptions. Unfortunately, in actual situations, the assumptions introduced are often violated due to measurement noises as well as irregular movements, such as suddenly turning the moving directions as well as interactive collision avoidance.

We have developed another type of measurements making use of occlusions caused by cluttered obstacles in office environments. Its basic idea is explained as follows. Let us consider a typical occlusion where a person occludes an obstacle or vise versa. Making use of the photometry, it is naturally viewed that the occlusion measurement provides some relationship on depth such that the person is further (or nearer) from the camera than the obstacle. Monitoring such depth information over time, we can even reason about the relationship on depth among multiple persons with respect to the same obstacle. We use this novel type of relative measurements together with the conventional absolute 3D measurements.

In this paper, we propose a novel technique for probabilistic visual tracking using both the absolute and the relative measurements in the particle filtering framework. Unlike absolute 3D measurements, relative measurements depend not only on the pose of a single person but on the relationship between poses of a pair of persons. This makes it difficult to directly apply standard techniques of particle filtering based on SIR sampling. We will show how to indirectly deal with the dependency using Markov Chain Monte Carlo -based Particle Filter (MCMC-PF) which has been originally proposed in [6] and recently attracting much interest. We demonstrate the effectiveness of propose techniques in a simple and difficult situation where two persons frequently occluding each other.

2. TRACKING METHOD

This section firstly introduces assumptions on the configuration on the camera as well as persons being tracked, then explains how to acquire the relative depth measurement between an obstacle and each person. It secondly provides a simple method for estimating the relationship on depth between a pair of persons using multiple measurements. It finally shows how to integrate both the relative depth measurements and the standard 3D absolute measurements within the probabilistic framework of particle filter.

2.1 The configuration of camera and persons

We consider typical settings of such as visual surveillance systems as well as intelligent spaces. The camera is fixed at a

location in an office higher than persons to be expected as shown in Figure 1. It can be seen that when two persons come near to each other their corresponding regions may occlude with one another in the image frame. This makes the identification of persons a difficult task.

We also introduce the following condition for regularizing the number of persons in the scene.

1. Two persons do not occupy the same 2D region in the real world space.
2. Events of the appearance and the disappearance of persons do not occur.

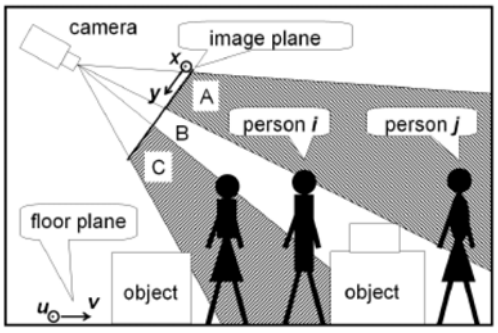
The condition #1 is a physical constraint. The condition #2 means that the persons do not appear nor disappear at any location except for boundaries of the scene. In other words, there is no occlusion where a person is completely occluded by some obstacles. Although such a complete occlusion is not dealt with in the current paper, there have been proposed some technique for counting the number of persons in terms of the occlusions such as [6].

We represent the pose of a person on the image and on the floor planes respectively as *x*-*y* and *u*-*v* coordinates as shown in Figure 1(a) where *v* is the depth direction in terms of the camera viewpoint.

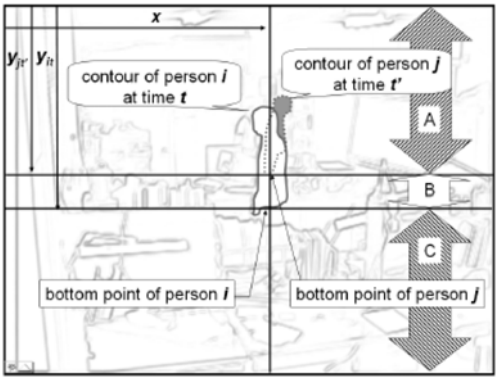
2.2 Acquisition of Relative Measurements

The region of a moving object in the scene is the region where the person occludes some static objects including background objects. In such a occlusion region, it is likely that the person is located in front of the static objects being occluded with respect to the camera viewpoint. So, it can be viewed that the measurement of an occlusion region contains some information on the information on depth of the moving object. More concretely, an occlusion region depends not only the depth but also other state such as shape of the moving object. So, pure depth feature has to be extracted from the occlusion region.

The general task of feature extraction is a difficult task and out of the scope of this paper. In this paper, we only deal with a case study in terms of the feature extraction. For our experimental system, we introduce a specific technique described in the following. Now, let us analyses the bottom boundary of the region of a moving object as shown in Figure 1(b). It can be seen that the bottom boundary is relatively invariant to shape of the moving object being observed, but is rather dependent only on the shape of the occluder obstacle in front of the person. It can be said that the bottom boundary is a depth feature weakly dependent on the shape of the target object. From the reason, we use each point on the bottom boundary (called a bottom point) as a feature point representing the relative depth measurement. More concretely, a bottom point is a point where the *y* value is largest for each *x* within the region of a moving object. The intersections of the vertical and the horizontal lines shown in Figure 1(b) indicate a pair of bottom points respectively corresponding to a pair of persons i and j at two different points *t* and *t'* in time. It is likely that a person is in front of those obstacle that are located above the bottom point on the image and also in back of those obstacles that are located below.



(a) position of the camera and persons



(b) Position measurements from bottom points

Figure 1: Relative position measurement.

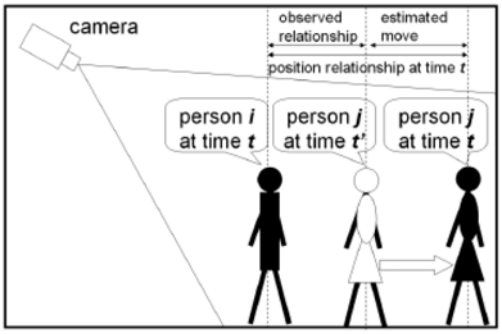


Figure 2: Estimation of position relationship.

2.3 Estimation of Depth Relationship

Now, we consider how to estimate the depth relationship from a pair of relative measurements on the same obstacle *o* on the image. We firstly consider a relatively simple situation where the two measurements arrive within a sufficiently short period in time. Let t and *t'(t>t')* denote the two points in time where the measurement arrive. Such a situation is described as the following conditions.

- The person *j* is likely to be in back of the obstacle *o* at *t'*.
- The person *i* is likely to be in front of the obstacle o at *t*.

In the above situation, it is likely that the person *i* is in front of the person *j* at *t*. In general, the above assumption that *t* and *t'* are near to each other is often violated. So, we need to consider also the movement of the target *j* between the time period *[t',t]*.

In the general case, it can be viewed that the target *i* is likely to be in front of the target *j* when the following two

conditions are satisfied.

1. The condition of person *i* at *t* being in front of person *j* at *t'*, which should be judged using the relative measurements.
2. The condition that person *j* moves in the depth direction with respect to the camera viewpoint, which should be judged using the absolute measurements.

In the following, we provide more concrete procedure of either condition.

The condition #1 is judged using bottom points extracted from relative measurements. Let (x, y_{it}) denote a bottom point of person *i* at *t*. From section 2.2, it can be seen that those obstacles that are located above the bottom point (x, y_{it}) are likely to be in front of person i, while those obstacles that are located below are likely to be in back of person *i*. In consequence, persons *i* at *t* is likely to be in front of person *j* at *t* (as shown in Figure 2) if a bottom point $(x, y_{jt'})$ of person *j* at *t'* satisfies

$$y_{jt'} < y_{it} - D_1 \quad (D_1 > 0) \tag{1}$$

as shown in Figure 1(b). This is because of that in such a case, all the obstacles located on a line segment (the line segment 'B' in Figure 1(a),(b)) are likely to be in back of person *i* and in front of person *j*. In the eqn (1), D_1 is a margin for the measurement noises on bottom points.

The condition #2 is judged using movements extracted from absolute measurements. Let v_{jt} denote the *v* element of the absolute measurement on a target object *j* at *t*. The movement in the *v*-direction between a time period *t', t (t'<t)* can then be represented as $v_{jt} - v_{jt'}$. When

$$v_{jt} - v_{jt'} > D_2 \quad (D_2 > 0), \tag{2}$$

it is likely that the *v*-locations of person *j* is larger at *t* than at *t'* in time as shown in Figure 2. Here, D_2 is a margin for measurement noises on absolute measurements.

When the above two conditions (1) and (2) are satisfied, it is likely that person *i* at *t* is in front of person *j* as shown in Figure 2. Naturally, the same thing holds for the case where the two persons are located in opposite relationship in depth. To enhance robustness, the two conditions are judged for each point on the bottom boundary and then the final decision is made based on whether over half the bottom points agree or not.

Direct implementation of the above idea requires all the history of bottom points as well as depth information

$$C_j[x] = (t', y_{jt'}, v_{jt'}) \quad (t' < t) \tag{3}$$

stored on the memory. The required space cost increases in an unbounded fashion as the number of bottom points arrive at each frame. In our current implementation, only the latest history is stored on the memory for each x coordinate on the image. More concretely, our method represents the history of person *j* by a 1D array

$$H_j = [C_j[1], \cdots, C_j[x_m]] \tag{4}$$

and overwrite each element every time a new bottom point arrives. This modified method requires only a constant time memory space whose size is proportional to the image width x_m.

2.4 Probabilistic Tracking

It has been described that the proposed relative measurements are dependent on the state of two persons. This is essentially different from conventional absolute measurements that are dependent only on the state of a single person. In this section, we will show how to integrate such different types of measurements in a consistent manner within a probabilistic framework of particle filtering.

The problem is formulated in a probabilistic framework. Let *I* denote the number of persons being tracked. $X_{it} = (u_{it}, v_{it})$ denote the pose of person $i(\in [1,I])$ at *t* on the floor. We associate with person *i* a object region on the image by which the likelihood $P(Z_{it}|X_{it})$ is maximized. Then, the measurement Z_{it} is represented using absolute measurement A_{it} and relative measurement C_{it} as

$$Z_{it} = (A_{it}, C_{it}) \tag{5}$$

Now, we represent the joint measurement and the joint state of all persons respectively as

$$Z_t = (Z_{1t}, \cdots, Z_{It}) \tag{6}$$

$$X_t = (X_{1t}, \cdots, X_{It}), \tag{7}$$

and the measurement sequence so far as

$$Z^t = (Z_1, \cdots, Z_t). \tag{8}$$

Then, the problem is formulated as estimating the probability density of the joint state being X_t conditioned on the measurement sequence Z^t. This density is computed in the following recursive form

$$P(X_t|Z^t) = P(Z_t|X_t) \cdot \int P(X_t|X_{t-1}) P(X_{t-1}|Z^{t-1}) dX_{t-1}. \tag{9}$$

In the formula, $P(X_t|X_{t-1})$ is called a joint motion model and describes how the joint state of persons moves over time, while $P(Z_t|X_t)$ is called a joint measurement model and describes the conditional density of a possible measurement arrives.

The probability density in (9) is a non-Gaussian and multi-modal distribution and dealt with by a particle filtering framework. A particle filter approximates the density $P(X_t|Z^t)$ as a set of discrete samples $S_t = \{X_t^{(n)} | 1 \leq n \leq N\}$ called particles. With this approximation, the density $P(X_t|Z^t)$ in (9) is computed in an incremental manner every time new motion and measurements arrive. It is known that the time cost of a particle filter is exponential to the dimensionality of the state space. This makes a naive implementation intractable in the high dimensional case. The problem of high dimensionality is efficiently dealt with approximately linear complexity in the framework of MCMC-PF. In the MCMC-PF framework, the joint motion and measurement models are represented by a set of independent potential functions for measurement $P(A_{it}|X_{it})$ and motion $P(X_{it}|X_{i,t-1})$ as well as models for interaction term $\phi(X_{it}, X_{jt})$ between a target pair in the form

$$P(X_t|Z^t) = \prod_i P(A_{it}|X_{it}) \prod_{ij} \psi(X_{it}, X_{jt}) \cdot \int \prod_i P(X_{it}|X_{i,t-1}) P(X_{t-1}|Z^{t-1}) dX_{t-1}. \quad (10)$$

The potential function $\phi(X_{it},X_{jt})$ takes a low value when the relationship between the states of the two targets are inadequate or a high value when otherwise.

In our scheme, the independent models as well as the potential functions are designed as follows. Considering that the absolute measurement A_{it} depends only on the state X_{it} of a person i at t, while the relative measurement C_{it} depends on the states X_{it}, X_{jt} of two persons, our measurement model is represented in the form

$$P(Z_t|X_t) = \prod_i P(A_{it}|X_{it}) \prod_{ij} P(C_{it}|X_{it}, X_{jt}). \quad (11)$$

The conditions for regularization introduced in section 2.1 are represented in the relationship $P(X_{it}, X_{jt})$ between two person *i*, *j* at *t*. With the conditions, our motion model is represented in the form

$$P(X_t|X_{t-1}) = \prod_i P(X_{it}|X_{i,t-1}) \prod_{ij} P(X_{it}, X_{jt}). \quad (12)$$

Putting the models (11) and (12) into (9), we obtain the eqn (10) where

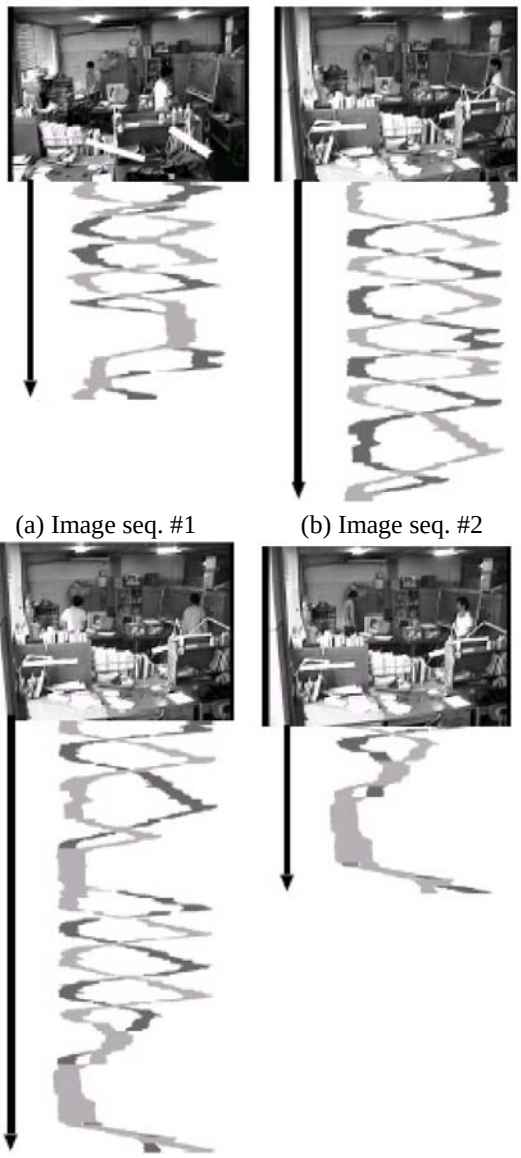
$$\psi(X_{it}, X_{jt}) = P(C_{it}|X_{it}, X_{jt}) P(X_{it}, X_{jt}). \quad (13)$$

2.5 Probabilistic Models

In current implementation, the measurement model (11) and the motion model (12) as well as the regularization conditions rely on several user defined parameters. In the following, α, β, γ are the normalization constants used by particle filtering and do not affect the final results.

The relative measurement model $P(C_{it}|X_{it}, X_{jt})$ is designed so that it takes α when the depth relationship between X_{it} and X_{jt} is consistent with the relative measurement C_{it} or α $exp(-G_H)$ otherwise. Here, G_H (>0) is a penalty value for errors in estimating the relationship between persons.

The absolute measurement model $P(A_{it}|X_{it})$ is designed in the similar manner as in the conventional methods. The probability density $P(A_{it}|X_{it})$ is given by the normal distribution of the error Δd between the predicted and the actual poses of the target being considered. The prediction is obtained by transforming the pose X_{it} of person from the floor coordinate to the image coordinate. For this, the head of person detected in the image is used as the feature point to represent the absolute pose of a person on the image. The use of head as a feature point is a major strategy used in many literatures such as in [3]. The head point is extracted from a set of points on the top boundary of the object region being considered as the median value of the top points. Here, each top point is defined for each *x* location in the object region as the point with smallest *y* value.



(a) Image seq. #1 (b) Image seq. #2

(c) Image seq. #3 (d) Image seq. #4

Figure 3 : Image sequences and tracking results.

The motion model $P(X_{it}|X_{i,t-1})$ is designed so that it takes a value β when the estimated velocity of person *i* on the floor is lower than a predefined threshold $D_3(>0)$ or $exp(-G_S)\beta$ otherwise. Here, $G_S(>0)$ is the penalty for velocity of move. The regularization model $P(X_{it},X_{jt})$ is designed so that it takes a value γ when the relationship between the states X_{it}, X_{jt} of two persons satisfy the regularization conditions (1),(2) described in section 2.1 otherwise a value γ multiplied by $exp(-G_H)$ or $exp(-G_O)$ depending on whether each condition is satisfied or not. Here, G_H,G_O are preset penalty values. More concretely, the penalty for condition #1 is assessed when the pose of two persons *i*, *j* overlap on the image and their states occupy the same region on the floor. The penalty for condition #2 is assessed when an object region in the image is associated with two different persons *i*, *j*.

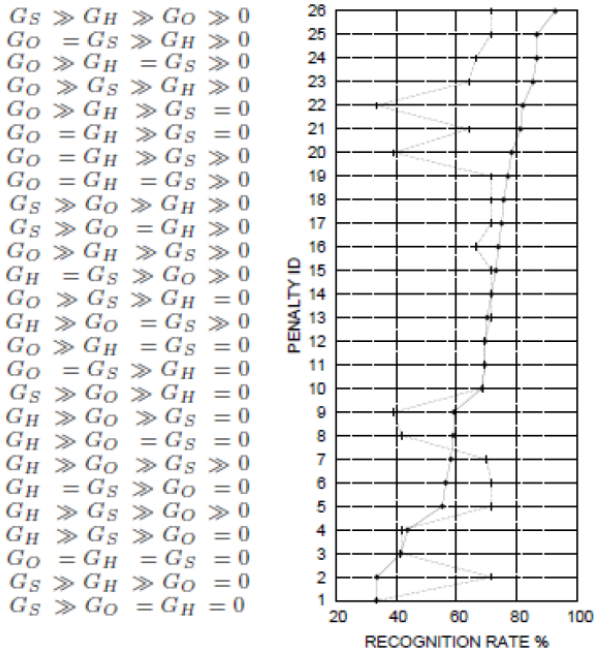
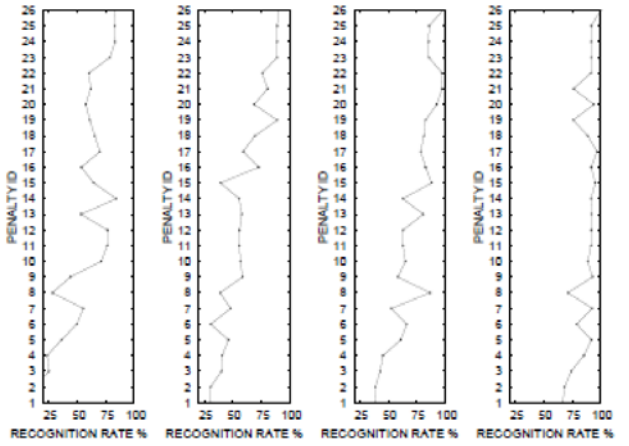
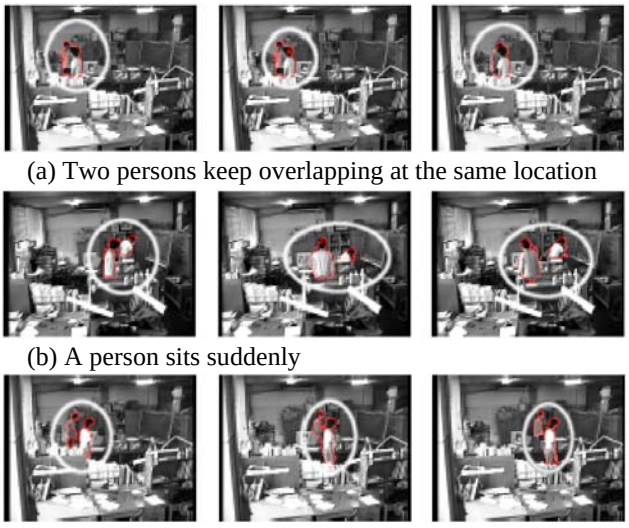


Figure 4: Performance comparison.



(a) Seq. #1 (b) Seq. #2 (c) Seq. #3 (d) Seq. #4

Figure 5: Results for each image sequence.



(a) Two persons keep overlapping at the same location

(b) A person sits suddenly

(c) Two persons keep overlapping while walking

Figure 6: Examples of critical situations.

3. Evaluation Experiments

This section provides experimental results to investigate the effectiveness of methods. We will compare the proposed method using both relative and absolute measurements against conventional method using absolute measurements alone. We also investigate the sensitivity of the proposed method against penalty values G_O, G_H,G_S shown in section 2.

3.1 Setups

The margins introduced in section 2.3 are empirically set *D1=10[pixel]*, *D2=2[m]*, while the threshold on movement velocity described in 2.5 is set *D3=5[m/s]*. A robust method for background subtraction similar with the one in [7] is implemented in our system. With the background subtraction method, there are negligibly small number of false negatives of object regions. However, there are much amount of false positives, which makes our recognition a difficult task. The ground truth data are manually created by indexing the input images with the person ids. With the ground truth, the estimation results will be evaluated in terms of recognition ratio by comparing the ground truth data with the region id output by the tracking method being tested. A recognition result is considered as correct when all the object regions are assigned correct person ids. The recognition ratio is defined as the ratio of the number of image frames where all the recognition results are correct. According to the definition, it might be argued that the recognition is always successful when the number of person regions is only 1. However, the correct case is equally relevant for every tracking method being tested. So, it does not cause serious problem from the viewpoint of performance comparison.

Figure 3 shows a set of 4 input image sequences to be used in this experiment. Each image is taken at image size 320 × 240 and frame rate 10[Hz]. The two line curves shown in each figure are trajectories of persons estimated by the vision systems. The left and the down directions in the figure correspond respectively to the *x*-location and the point in time of person being considered. Note that the intersection points in the figure indicate that the persons overlap in the same image frame. They are situations where the identification of persons becomes a difficult task. In many cases, such situations are classified into one of the following cases.

1. Multiple persons come close to each other on the image plane.
2. Multiple persons come close to each other on the floor plane.
3. Multiple persons are moving on the same path.

In the current experiments, the path in the case #3 is realized as a loop path surrounding several obstacles including tables and desks.

3.2 Results

Figure 4 shows the result of comparison between the proposed and the conventional methods. In the experiments, a number of recognition tasks have been conducted for the entire image sequence as well as individual image sub-sequences and for various settings of methods and

penalty values. Each result is evaluated in terms of the recognition ratio. For the sake of clarity, the results are sorted in ascending order of the recognition ratio against the entire image sequence, then based on the scoring results, a unique penalty id is assigned for each setting of penalty values. 'o' and '+' are the recognition ratio of proposed and conventional methods for the entire image sequence.

It can be seen from the figures that the proposed method yields better results than the conventional method. Especially, the recognition ratio of the proposed method exceeds 90% for some combinations of parameter settings. The conventional method does not exceeds 70%. On the other hand, the combinations where the recognition ratio of the proposed method is worse than that of the conventional method is the penalty ids = 2,4,5,6,7,13. Note that these settings can be explained by an inequality $G_H >> G_O$ which means that it puts more weight on the penalty G_O for the appearance / disappearance of persons than the penalty G_H for the history of relative measurements. With such a penalty setting, it becomes difficult to recover from accumulated errors in movement when relative measurements are unreliable. On the contrary, high recognition ratio with over 74% is obtained for those settings (penalty ID = 16,18,22,23,24,25) summarized by an inequality

$$G_O \gg G_H \gg 0. \qquad (14)$$

From above results, it can be concluded that the proposed method is effective for improving the recognition ratio of a tracking system.

Figure 5 shows results by the proposed method for the 4 individual image sub-sequences shown in Fig. 3(a), (b), (c), (d). The sub-sequence #1 contains a difficult situation where the two persons being tracked overlap to each other on the image for a long period of time as shown in Fig. 6(a). Moreover, one of the persons suddenly sat down just before the two persons intersect as shown in Fig. 6(b). Such an event naturally cause large amount of errors in absolute pose measurements. Due to such events, the recognition performance is not sufficiently high as shown in Figure 5 (a). The sub-sequence #2 illustrated in the Figure contains only a simple intersection within a short period of time. In that case, the recognition ratio is high 72% for almost all penalty values that satisfy inequality (14), except for penalty ID = 18 where the recognition ratio is 69.5%. The sub-sequence #3 illustrated in Figure 6(c) contains a difficult situation where two persons walk in the same direction on the image while overlapping to each other. In the case, a difficulty arises from the fact that no relative measurement arrive during the events. In spite of that, the proposed method yields high recognition ratio over 81% for every combination of penalty values that satisfy (14). From the results, it can be viewed that relative measurements are robust against temporary noises. The sub-sequence #4 also contains the difficult situation shown in Figure 6(c). In spite of that high recognition ratio over 88% is obtained for those penalty values that satisfy (14).

4. CONCLUSIONS

This paper introduced a novel scheme for vision-based multiple person tracking. This scheme makes use of relative pose measurements from occlusion reasoning as well as conventional absolute measurements. This makes the proposed scheme robust against complex occlusion situations such as two persons overlap or occluded each other. This paper also presented how to integrate the different types of measurements as well as their uncertainties within the probabilistic framework of particle filtering. Likewise, we plan to integrate other types of measurements such as appearance to enhance the robustness of our visual tracking system.

ACKNOWLEDGEMENT

This work was partially supported by MECSST Grant in-Aid for Young Scientists (B) (17700200) and by Suzuki Foundation Research Grant. The authors are grateful for useful advice from Dr. Y. Kimuro, N. Okada and E. Kondo.

REFERENCES

[1] Isard M. and Blake A. Condensation – conditional density propagation for visual tracking. *Int. J. Computer Vision*, 29(1):5–28, 1998.

[2] Arulampalam M. S., Maskell S., Gordon N., and Clapp T. A tutorial on particle filters for online non-linear/non-Gaussian Bayesian tracking. *IEEE trans. signal processing*, 50(2):174–188, 2002.

[3] Rosales R. and Sclaroff S. Improved tracking of multiple humans with trajectory prediction and occlusion modeling. *CVPR Workshop on the Interpretation of Visual Motion*, 1998.

[4] Schult D., Burgard W., Fox D., and Cremers A. B. People tracking with mobile robots using sample-based joint probabilistic data association filters. *Int. J. Robotics Research*, 22(2):99–116, 2003.

[5] James W. Davis and Stephanie R. Taylor. Analysis and recognition of walking movements. *Proc. Int. Conf. Pattern Recognition*, pages 315–318, 2002.

[6] Zia Khan, Tucker Balch, and Frank Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(11):1805–1918, 2005.

[7] Satoh Y., Tanahashi H., Wang C., and Kaneko S. Robust event detection by radial reach filter (rrf). *Proc. IEEE Int. Conf. Pattern Recognition*, pages 623–627, 2002.